Dual Branch Cooperative Learning for Domain Adaptive Deepfake Detection

Anonymous submission

Abstract

Recent advancements in deepfake detection have shown promising results when training (source domain) and testing (target domain) involve face forgeries from the same dataset. However, challenges arise when attempting to generalize the detector to detect forgeries created using unfamiliar techniques not present in the training dataset. This work addresses the generalizable deepfake detection from a simple principle: a domain-invariant representation should be robust against diverse types of forgeries. Following this principle, we propose a novel Dual-Branch Collaborative Learning (DBCL) framework. First, we generate latent domains using fast Fourier transform, and based on this, we construct a new adaptation space consisting of multiple pairs (i.e. “source→target” and “latent→target”), enabling DBCL to perform image-level and feature-level alignment in a unified perspective. Second, our DBCL assigns high-quality pseudo-labels to the target domain through knowledge distillation and employs a dynamic updating module to refine these labels, thereby mitigating the adverse impact of pseudo-label noise on domain adaptation. Extensive experiments show that our DBCL performs better than the state-of-the-art methods, demonstrating its effectiveness for domain adaptive deepfake detection.

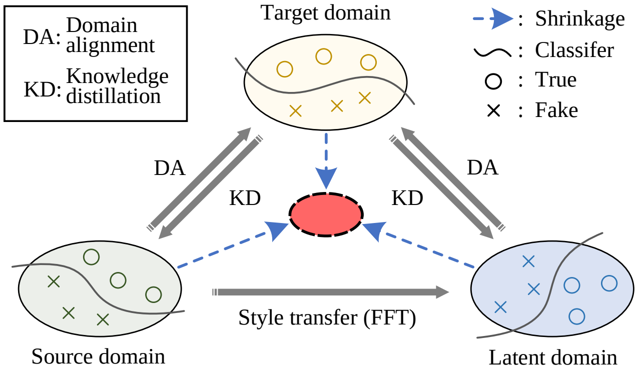

Figure 1: Core idea of our DBCL framework. We generate latent domain data online through FFT, and based on this, construct two domain adaptation pairs, i.e. “source→target” and “latent→target”. In each pair, domain alignment and knowledge distillation mutually promote each other to effectively reduce domain discrepancies. Through our DBCL, three domains tend to be shrunk into a novel compact distribution (as illustrated by the red ellipse), yielding discriminative domain-invariant representations.

Introduction

With the rapid advancement of deep learning technology, facial forgery techniques (a.k.a. deepfake) (Hong et al. 2022; Mazaheri and Roy-Chowdhury 2022; Xu et al. 2022; Zhai et al. 2022; Mittal et al. 2023) have become increasingly widespread, posing profound impacts on both society and individuals. Deepfake technique can be exploited for disseminating false information, identity impersonation, and fraudulent activities, significantly undermining information credibility and personal privacy. Therefore, the development of effective deepfake detection methods is of great importance.

The deepfake detection technology can be streamlined into a binary classification problem, aiming to differentiate whether the face is authentic or has been forged. With the powerful learning capability of CNNs, deep learning-based methods for deepfake detection have made significant progress (Nirkin et al. 2022; Chen et al. 2021b; Jeong et al. 2022; Li et al. 2021b; Haliassos et al. 2022). However, most methods primarily focus on enhancing detection accuracy within individual datasets, without considering the generalization ability of the model. With the rapid advancement of deepfake technology, the variety of face forgery methods becomes increasingly diverse, leading to numerous datasets (Rössler et al. 2019; Li et al. 2020b; He et al. 2021) containing various types of forgery. When training data (source domain) and testing data (target domain) originate from different deepfake techniques, the model trained only on the source domain exhibits a pronounced decline in detection accuracy when apply to the target domain. Deepfake detection methods tailored for specific types of forgeries are no longer sufficient to meet practical demands. For this purpose, researchers have focused on how to improve the generalizability of deepfake detection methods.

In recent years, Unsupervised Domain Adaptation (UDA) techniques (Ye et al. 2022; Mirza et al. 2022; Gao et al. 2022) have gradually emerged as the dominant approach for enhancing model generalization, aiming to transfer models from the source domain to the target domain by learning domain-invariant representations. Most UDA methods (Yu et al. 2022; Zhang et al. 2022b; Li et al. 2021a) reduce the discrepancies between the distributions of source and target domains either at the image level or the feature level. There are also some methods (Lee et al. 2022; Manjah et al. 2023; Huang et al. 2023b) that assign pseudo-labels to the target domain using self-training based knowledge distillation techniques, alleviating the difficulty of domain adaptation. Despite significant progress having been made, existing domain adaptive deepfake detection methods still encounter issues that constrain their generalization performance: **(1) How to efficiently align domains?** To achieve domain alignment at the image level, some methods require training independent style transfer models to generate latent domains, which reduces the difficulty of domain adaptation, but introduces a significant amount of additional computation. Moreover, the existing UDA methods employ a single pair framework ("source→target" or "latent→target") to facilitate adaptation, without considering the extraction of richer cross-domain knowledge from the joint distribution space. **(2) How to distill high-quality knowledge?** While utilizing knowledge distillation techniques to generate pseudo-labels, models can be more effectively transferred to unlabeled unknown target domains. However, the significant visual differences between the source and target domains do affect the quality of these pseudo-labels. The potential for mislabeling within the pseudo-labels can readily trigger error propagation, consequently detrimentally influencing the model's domain adaptation process.

To this end, we propose a Dual-Branch Collaborative Learning (DBCL) framework to learn domain-invariant representations for domain adaptive deepfake detection. To achieve more effective domain alignment, we utilize FFT to dynamically generate target-like latent images, and based on this, we construct a new adaptation pipeline that incorporates multiple pairs, enabling effective domain alignment at both the feature and image levels. To distill high-quality knowledge for the target domain, our DBCL employs a teacher-student network with two branches. In this setup, the student network enhances the quality of pseudo-labels generated by the teacher network through domain alignment across multiple paired domains. Simultaneously, the teacher network employs a dynamic updating strategy to provide reliable pseudo-labels to the student network, thereby reducing potential erroneous noise within the pseudo-labels. As illustrated in Figure 1, our DBCL performs domain alignment and knowledge distillation in a mutually reinforcing manner within each pair (i.e. "source→target" and "latent→target"), ultimately aggregating the distributions of the three domains to generate discriminative domain-invariant features. We summarize contributions below:

- We propose a novel domain adaptive deepfake detection method, named DBCL. To the best of our knowledge, this is the first endeavor for deepfake detection through domain alignment and knowledge distillation in a mutually reinforcing manner.
- Our DBCL framework generates high-quality pseudo-labels through multi-level domain alignment. Moreover, we introduce a dynamic updating strategy to refine these pseudo-labels, effectively mitigating the adverse impact of noise on the transfer model.
- Experiments on benchmark datasets demonstrate that our DBCL achieves a new state-of-the-art performance. Ablation study also validates the effectiveness of each component in DBCL for domain adaptive deepfake detection.

Related Work

Unsupervised Domain Adaptation. For an unlabeled target dataset, UDA aims to enhance the model's deepfake detection accuracy on the target domain by aligning the source and target domains. Common methods for domain alignment mainly include image-level and feature-level alignment. *(1) Image-level alignment.* These methods are primarily achieved through techniques such as style transfer (Yang and Soatto 2020; Kundu et al. 2022), data augmentation (Araslanov and Roth 2021; Melas-Kyriazi and Manrai 2021), and image mixing (Wu et al. 2022). Most image-level alignment requires training specialized models to generate latent domain data, which reduces the difficulty of alignment but adds extra computational overhead. *(2) Feature-level alignment.* This kind of method aims to reduce distribution shifts between two domains. Researchers commonly utilize the Maximum Mean Difference (MMD) (Kumagai and Iwata 2019; Chen et al. 2020) and its variants as discrepancy measures. Another approach involves using a domain discriminator to reduce the domain discrepancy (Zhu et al. 2022; Akkaya, Altinel, and Halici 2021). In this study, we employ FFT to dynamically generate latent domains in a more flexible manner. Subsequently, we construct a new framework involving two pairs of adaptations, which unifies image-level and feature-level alignment.

Pseudo-Label Based Self-Training. To assign high-quality pseudo-labels to unlabeled target domain data, self-training (Zheng et al. 2023; Yu et al. 2021) can be used to distill knowledge from both source and target domains for the target domain. A core concept in self-training is teacher-student optimization, closely related to knowledge distillation. The teacher network generates pseudo-labels for unlabeled target domain data, guiding the student network to adapt better to the target domain. When knowledge is distilled from an advanced teacher network to a student network, it improves the performance of the student network. The quality of pseudo-labels significantly impacts the performance of the student network. When pseudo-labels contain noise (i.e. incorrect categories), it adversely affects model optimization. Hence, some research focuses on improving the quality of pseudo-labels(Lai et al. 2022; Petrovai and Nedevschi 2022; Cheng et al. 2023; Li et al. 2022), while others concentrate on reducing noise to help the model optimization (Liu et al. 2021c; Xu et al. 2023; Chen et al. 2021a; Qu et al. 2023; Nishi et al. 2021). When there's a significant domain discrepancy between the source and target domains, the generated pseudo-labels become unreliable. In this study, we employ domain alignment to reduce the gap between source and target domains, facilitating knowledge distillation for generating high-quality pseudo-labels. Additionally, we propose a dynamic updating strategy to mitigate the adverse effects of erroneous pseudo-labels.

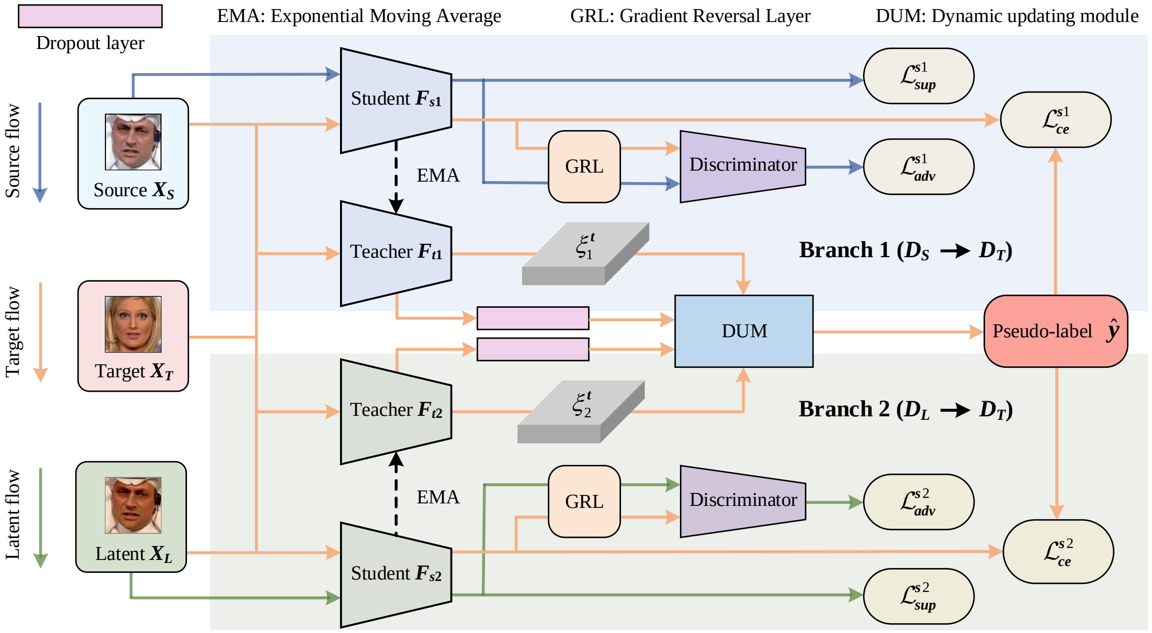

Figure 2: Overview of our Dual-Branch Collaborative Learning (DBCL) framework. The proposed DBCL framework consists of two branches, with each branch comprising a teacher and a student network. The two branches are applied to the pairs of "$D_S \rightarrow D_T$" and "$D_L \rightarrow D_T$", respectively. In each branch, the student network is employed for domain alignment, facilitating the knowledge distillation from the corresponding teacher network. To refine $\hat{y}$ generated by F_{t1} and F_{t2}, we introduce a Dynamic Updating Module (DUM), mitigating the adverse impact of pseudo-label noise on the optimization of F_{s1} and F_{s2}.

Domain Adaptive Deepfake Detection. Most current deepfake detection methods rely on deep neural networks, employing primary techniques such as X-ray detection (Li et al. 2020a), lip motion detection (Haliassos et al. 2021), frequency domain detection (Liu et al. 2021a; Luo et al. 2021), along with other artifact-independent methods (Yeh et al. 2021; Wang et al. 2022; Huang et al. 2022). With the rapid development of deepfake technology, the diversity of forgery methods results in domain discrepancies among different forgery images. How to improve the generalization capability of deepfake detection has become the mainstream research in this field. Current generalized deepfake detection techniques primarily focus on three aspects, i.e. data perspective (Shiohara and Yamasaki 2022; Zhang et al. 2022a; Chen et al. 2022), feature perspective (Zheng et al. 2021; Luo et al. 2021; Fei et al. 2022), and learning strategy perspective (Kim, Tariq, and Woo 2021; Dong et al. 2022). In this study, we propose a novel domain adaptive deepfake detection method, which simultaneously performs domain alignment and knowledge distillation within a unified framework. By mutually promoting each other, the detection ability of the model in the target domain is enhanced.

Method

Overview

For domain adaptive deepfake detection, we propose a Dual-Branch Collaborative Learning (DBCL) framework that can transfer models trained in the labeled source domain to the unlabeled target domain. DBCL consists of two branches, each containing a teacher network and a student network, as illustrated in Figure 2. The teacher and student of the first branch are named F_{t1} and F_{s1}, respectively, while those of the second branch are named F_{t2} and F_{s2}. To achieve more flexible image-level alignment, we use an FFT-based style transfer method (Yang and Soatto 2020) to generate images X_L in the latent domain, preserving the content of the images X_S in the source domain with the style of the images X_T in the target domain. Specifically, we use FFT to generate target-like images X_L by replacing the amplitude of X_S with that of X_T.

For domain adaptive deepfake detection, we define the source domain, latent domain, and target domain as $D_S{=}\{(X^i_S, Y^i_S)\}^{N_S}_{i=1}$, $D_L{=}\{(X^j_L, Y^j_S)\}^{N_L}_{j=1}$, and $D_T{=}\{X^k_T\}^{N_T}_{k=1}$, respectively, where N_S, N_L, and N_T denote the number of images in each domain. Since X_S and X_L share the same content, Y_S can serve as the label for them simultaneously. We construct two pairs of adaptation, i.e. "$D_S \rightarrow D_T$" and "$D_L \rightarrow D_T$", where both pairs undergo joint domain alignment and knowledge distillation, mutually reinforcing each other to extract domain-invariant representations. Specifically, F_{t1} and F_{s1} from the first branch are used for domain alignment and knowledge distillation on the "$D_S \rightarrow D_T$" pair, while F_{t2} and F_{s2} from the second branch handle the same processes within the "$D_L \rightarrow D_T$" pair.

In each branch, we utilize the Adam optimizer to update the parameters of the F_s and then use these updated parameters to update the parameters of F_t through Exponential Moving Average (EMA) (Tarvainen and Valpola 2017). F_{s1}

takes images from the D_S and D_T as input, while F_{s2} takes images from the D_L and D_T as input. F_{s1} and F_{s2} perform feature alignment through adversarial learning, each generating domain-insensitive features with respect to the D_T and another domain, respectively. The X_T simultaneously serves as input to both F_{t1} and F_{t2}, distilling more comprehensive knowledge (i.e. pseudo-labels) from diverse domain pairs, which are utilized as supervisory labels for their corresponding F_{s1} and F_{s2}. Moreover, to alleviate the negative impact of pseudo-label noise on the transferred model, we introduce a Dynamic Updating Module (DUM). By optimizing the teacher-student networks in the two branches, domain alignment and knowledge distillation can mutually reinforce each other, promoting the generation of more robust domain-invariant representations. It's worth noting that in the inference stage, only F_{s1} and F_{s2} are used to predict the detection results for X_T.

Dual-Branch Cooperative Learning

In our DBCL framework, two student networks, F_{s1} and F_{s2}, are respectively utilized for domain alignment in the "source→target" and "latent→target" pairs. This alignment facilitates the distillation of precise knowledge from the corresponding teacher network, thereby assigning high-quality pseudo-labels to target images X_T. To simplify the explanation, we will use the "source→target" pair as an example to illustrate the optimization process of F_{s1} and F_{t1} in the first branch. The optimization principle in the second branch is the same as that in the first branch.

For the optimization of the student network, the domain alignment of the student network can facilitate the knowledge distillation of the teacher network. The student network can effectively align the D_T and other domains, which determines the teacher network's ability to distill high-quality knowledge. The significant disparities between D_S and D_T make domain alignment challenging. To address this, we first introduce an additional latent domain through FFT, which helps mitigate domain adaptation difficulties at the image level. Furthermore, adversarial learning is employed to achieve domain alignment at the feature level. By aligning the D_T and other domains separately in two branches, it allows the two branches to explore more abundant cross-domain knowledge through two student networks, providing more clues for the refinement of pseudo-labels. In F_{s1}, the inputs X_S and X_T pass through the feature extractor, the adversarial discriminator, and the domain classifier sequentially. Through adversarial learning, the model becomes incapable of distinguishing whether the feature extractor's output features originate from D_S or D_T, thereby acquiring domain-invariant representations. To train the student network in an end-to-end manner, we introduce a Gradient Reversal Layer (GRL) (Ganin and Lempitsky 2015) between the feature extractor and the domain classifier. When X_T in D_T lacks labels, the loss function of F_{s1} can be considered as a UDA loss, expressed as follows:

$$\mathcal{L}_{uda}^{s1} = \mathcal{L}_{sup}^{s1}(X_S, Y_S) + \lambda_{adv}\mathcal{L}_{adv}^{s1}(X_S, X_T, Y_d), \quad (1)$$

where $Y_d{=}\{0,1\}$ denotes the domain labels, λ_{adv} is the weight to balance the importance of two components in UDA loss, $\mathcal{L}_{sup}^{s1}$ and $\mathcal{L}_{adv}^{s1}$ denote the negative log-likelihood loss used for discriminating between real and fake faces and classifying features from which domain, respectively.

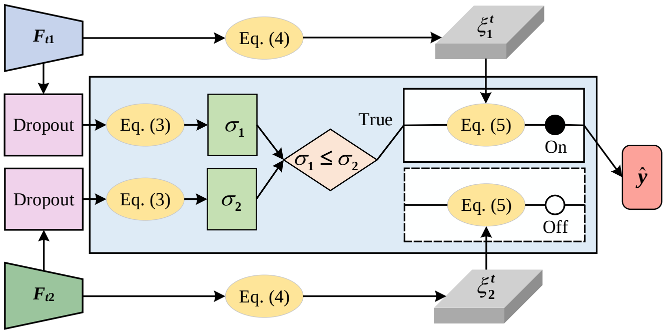

Figure 3: Illustration of our Dynamic Updating Module (DUM). We predict the multiple results of F_{s1} and F_{s2} through two dropout layers, and then calculate σ_1 and σ_2 using Eq. (3). When $\sigma_1{\le}\sigma_2$, we feed the ξ_1^t adjusted by Eq. (4) into Eq. (5) to generate $\hat{y}$. Conversely, when $\sigma_1{>}\sigma_2$, we feed ξ_2^t into Eq. (5) to generate $\hat{y}$. By utilizing DUM, we can obtain reliable pseudo-labels.

For the optimization of the teacher network, we utilize EMA to propagate the parameters from the student network to the teacher network, which can be represented as follows:

$$\theta_{t1} = \alpha\theta_{t1} + (1-\alpha)\theta_{s1}, \quad (2)$$

where θ_{t1} and θ_{s1} denote the parameters of the F_{t1} and F_{s1}, respectively, α denotes the smoothing coefficient of EMA, which determines the weight for updating the parameters of F_{t1}. For X_T, F_{t1} will predict whether it is true or fake, and use the prediction result as the pseudo-label for X_T.

Pseudo-Label Refinement for Self-Training

The pseudo-labels Y_{t1} and Y_{t2} of X_T can be predicted simultaneously through F_{t1} and F_{t2}. During the initial optimization stages of F_{s1} and F_{s2}, since the significant disparity between X_T and images in other domains cannot be reduced effectively, the credibility of Y_{t1} and Y_{t2} cannot be guaranteed. Due to the differences in domain alignment difficulty between F_{s1} and F_{s2}, the predicted results of F_{t1} and F_{t2} on facial images from the same target domain often exhibit ambiguity (i.e. $Y_{t1}{\neq}Y_{t2}$).

To enhance the reliability of the teacher network in generating pseudo-labels, we propose a Dynamic Updating Module (DUM) to refine the pseudo-labels, as shown in Figure 3. For the same X_T, we introduce two dropout layers to generate n predictions from F_{t1} and F_{t2} respectively, and then calculate the standard deviation σ of the n prediction results, which can be expressed as:

$$\sigma_i = \sqrt{\frac{\sum_{j=1}^{n}(T_i^j(X_T) - \mu)}{n}}, \quad (3)$$

where μ is the average of n prediction results, and $T_i^j(\cdot)$ denotes the j-th prediction result of F_{ti}. Inspired by (Li et al. 2022), to prevent the teacher network from forgetting history knowledge, we introduce an online adjustment strategy. This strategy updates the binary probabilities $H_i^p(\cdot)$ predicted by F_{ti} after the p-th epoch, which can be represented as:

$$\xi_i^t = \frac{p}{1+p} H_i^{p-1}(X_T) + \frac{1}{1+p} H_i^p(X_T), \tag{4}$$

where $H_i^{p-1}(\cdot)$ denotes the binary probabilities predicted by F_{ti} after the $(p-1)$-th epoch, and ξ_i^t represents the updated binary probabilities (i.e. the probabilities of being classified as a real or fake face).

According to Eqs. (3) and (4), we introduce a DUM to refine pseudo-labels. By comparing the values of σ_1 and σ_2, we choose the teacher network with lower uncertainty to generate pseudo-labels. The refined pseudo-labels for X_T can be obtained using the following formula:

$$\hat{y} = \begin{cases} \arg\max[\frac{\exp(\xi_1^t(h))}{\sum\limits_{z=1}^{C}\exp(\xi_1^t(z))}], \ \sigma_1 \leq \sigma_2 \\ \arg\max[\frac{\exp(\xi_2^t(h))}{\sum\limits_{z=1}^{C}\exp(\xi_2^t(z))}], \ \sigma_1 > \sigma_2 \end{cases}, \tag{5}$$

where C=2 is the number of categories, $\xi_i^t(h)$ denotes the predicted probability value in the h-th channel (h=1, 2).

Considering that pseudo-labels may still be incorrect, we use a smoothing labeling strategy to prevent the student network from overly relying on pseudo-labels, so that pseudo-labels can be better utilized. The predicted probability values ξ_i^s of F_{si} (i=1, 2) can be adjusted as follows:

$$\xi_i^s = \xi_i^s(1-\varepsilon) + \varepsilon/C, \tag{6}$$

where ε denotes the smoothing factor. When X_T obtains refined pseudo-labels $\hat{y}$, we use a cross-entropy loss $\mathcal{L}_{ce}^{si}$ to optimize F_{si} as:

$$\mathcal{L}_{ce}^{si} = -\sum_{z=1}^{C} \hat{y}_z \log(\xi_i^s(z)), \ \text{where } i = 1, 2. \tag{7}$$

Therefore, UDA loss in Eq. (1) can be transformed into a semi-supervised domain adaptation loss. Thus, the total loss function of F_{s1} and F_{s2} can be represented as:

$$\mathcal{L}_{semi}^{total} = \sum_{i=1}^{2} \left(\mathcal{L}_{uda}^{si} + \lambda_{ce}\mathcal{L}_{ce}^{si}\right), \tag{8}$$

where λ_{ce} is a hyperparameter used to control the importance of $\mathcal{L}_{ce}^{si}$. As illustrated in Algorithm 1, we obtain F_{s1} and F_{s2} with optimized $\hat{\theta}_{s1}$ and $\hat{\theta}_{s2}$, respectively. During inference, we perform the DUM on F_{s1} and F_{s2} to obtain reliable predictions for testing images in target domain.

Experiments

Experimental Settings

Algorithm 1: Dual Branch Cooperative Learning

Input: Source domain D_S and target domain D_T
Initialise: θ_{t1}, θ_{s1}, θ_{t2}, θ_{s2}, λ_{adv}, λ_{ce}, α, ε, n, $Maxiter$
Output: Optimized F_{s1} with $\hat{\theta}_{s1}$ and F_{s2} with $\hat{\theta}_{s2}$

```
for $iter = 1, 2, ..., Maxiter$ do
    Generate $D_L$ using FFT
    Update $\theta_{s1}$ and $\theta_{s2}$ by calculating $\mathcal{L}_{uda}^{s1}$ and $\mathcal{L}_{uda}^{s2}$ using Eq. (1)
    Update $\theta_{t1}$ and $\theta_{t2}$ using Eq. (2)
    for $j = 1, 2, ..., n$ do
        Calculate $\sigma_1$ and $\sigma_2$ using Eq. (3)
    end for
    if $\sigma_1 \leq \sigma_2$ then $i = 1$ else $i = 2$ end if
    Calculate $\xi_i^t$ using Eq. (4)   // Online Adjustment
    Generate $\hat{y}$ through $F_{ti}$ using Eq. (5)
    Adjust $\xi_i^s$ using Eq.(6)   // Smooth Labeling
    Update $\hat{\theta}_{s1} \leftarrow \theta_{s1}$ and $\hat{\theta}_{s2} \leftarrow \theta_{s2}$ by calculating $\mathcal{L}_{semi}^{total}$
        using Eq. (8), where $\mathcal{L}_{ce}^{si}$ is calculated using Eq. (7)
end for
```

Datasets. We conduct experiments on three benchmark datasets, i.e. Faceforensics++ (FF++) (Rössler et al. 2019), CelebDF (Li et al. 2020b), and ForgeryNet (He et al. 2021). **(1) FF++ dataset** consists of a total of 1,000 real videos (720 videos for training, 140 videos for validation, and 140 videos for testing) and 4,000 fake videos generated from the real videos using four different forgery methods, i.e. FaceSwap (FS), Face2Face (F2F), DeepFake (DF), and NeuralTextures (NT). Besides, the raw videos are compressed by H.264 to two levels, high-quality videos (HQ) and low-quality videos (LQ). **(2) Celeb-DF-v2 dataset** provides 590 real videos and 5,639 fake videos. The fake facial videos are generated by improved deepfake synthesis techniques, where those fake faces are more realistic and thus harder to distinguish for human eyes. **(3) ForgeryNet dataset** includes image-level and video-level data across four tasks: image forgery classification, spatial forgery localization, video forgery classification, and temporal forgery localization. In this work, we focus on image forgery classification, which contains 2,351,305 images for training and 158,201 images for evaluation. The fake images are generated by 15 forgery methods.

Implement details. For FF++, we follow the standard setting (Rössler et al. 2019), randomly selecting 100 frames from each video and employing Dlib (King 2009) to extract faces from images. For Celeb-DF-v2, we extract face images from 5,711 videos for training and from 518 videos for testing. For ForgeryNet, we use the image subset and produce facial image patches by applying a face detector RetinaFace (Deng et al. 2020). To prevent overfitting, we apply random data augmentations (i.e. horizontal flipping, rotation, adding Gaussian noise, color transformations, and erasure enhancement (Ni et al. 2022)) to the training data. In our DBCL framework, we employ Xception (Chollet 2017) as the backbone network for F_{t1}, F_{t2}, F_{s1}, and F_{s2}. The parameters θ_{t1}, θ_{t2}, θ_{s1}, and θ_{s2} are initialized using an Xception model pretrained on ImageNet (He et al. 2016). For the hyperparameters, we empirically set α=0.994, n=5, λ_{ce}=1.0, λ_{adv}=0.1, and ε=0.1. We implement our DBCL framework in PyTorch. The Adam optimizer with a base learning rate of 0.0002 is used to optimize F_{s1} and F_{s2}. All models are trained on 3 Tesla V100s. During the training stage, it takes approximately 9 hours (30k iterations) for the model to converge. During the inference stage, the processing speed averages around 4.4 frames per second (FPS).

Evaluation metrics & comparison methods. Following the standard setting (Rössler et al. 2019; Haliassos et al. 2021), we use area under receiver (AUC) as the evaluation metric. The AUC (%) shows the separability of real and fake faces. To demonstrate the effectiveness of our approach, we compare it with several state-of-the-art methods: FDA (Yang and Soatto 2020), LipForensics (Haliassos et al. 2021), CORE (Ni et al. 2022), UIA-ViT (Zhuang et al. 2022), DCL (Sun et al. 2022), and IID (Huang et al. 2023a).

Ablation Study

We validate the effectiveness of the key components of the proposed DBCL method from three perspectives.

(1) Effectiveness of the online adjustment strategy. As shown in Eq. (4), we introduce an updating approach to refine pseudo-labels, preventing the loss of historical knowledge distilled from the teacher network. To validate its effectiveness, we conduct experiments on the "NT→F2F" pair, where NT and F2F are the source domain and the target domain, respectively. For the predictions of the teacher network, we evaluate the accuracy of pseudo-labels for training and testing images in F2F with and without the online adjustment method. The results of the AUC values (%) are shown in Table 1. We can observe that when using our online adjustment method, the accuracy of pseudo-labels in the target domain significantly improves. We also find that there is only a 1.46% improvement in the training data, yet it leads to a 5.37% improvement in the testing data. This is due to that our OA strategy distills knowledge from the training set. Although it provides little assistance to the training set itself, it significantly aids in the model's adaptation to the testing set.

Method	Training set	Testing set
w/o OA	89.02	85.89
w/ OA	90.48	91.26
Δ	+1.46	+5.37

Table 1: Study on the online adjustment (OA) strategy. AUC values (%) are evaluated on the training and testing of F2F dataset. Δ is the performance gain of "w/" relative to "w/o".

(2) Effectiveness of the smooth labeling strategy. As illustrated in Eq. (6), we introduce a labeling strategy to better utilize pseudo-labels. To validate its effectiveness, we conduct experiments on pairs from different source and target domains, as shown in Table 2. Additionally, we explore the detection performance of the target domain under different smoothing factors in Eq. (6). As depicted in Figure 4, when the smoothing factor is set to 0.1, superior performance is achieved on all three domain adaptation pairs.

Source	Target	w/o SL	w/ SL
FS	F2F	86.62	**87.78**
DF	NT	86.53	**87.32**
NT	F2F	89.54	**91.26**

Table 2: Study on the smooth labeling (SL) strategy.

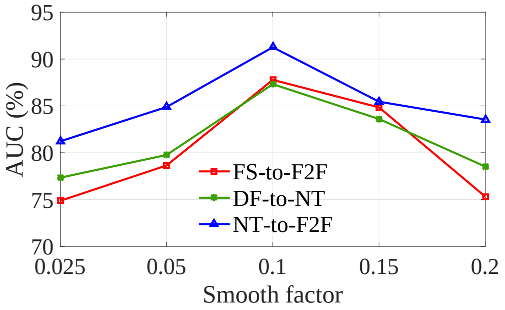

Figure 4: Study on the selection of the smoothing factor ε.

$D_S \rightarrow D_T$	①	②	③	④	⑤	⑥
FS→F2F	64.84	65.88	73.58	75.21	76.67	**87.78**
DF→NT	68.27	58.37	74.17	62.85	76.73	**87.32**
NT→F2F	69.88	67.81	72.35	77.28	81.38	**91.26**

Table 3: Ablation analysis on the optimization of different teacher-student frameworks. **Single-branch framework** concludes: ① Using the Xception network as the backbone for the student, and optimizing it solely with source data. ② Leveraging latent domain to optimize the student network through self-training. ③ and ④ Utilizing the source domain and latent domain to optimize the student network through adversarial learning, respectively. **Two-branch framework** encompasses: ⑤ Employing a consistency constraint strategy to generate pseudo-labels for the target domain. ⑥ Our proposed DBCL framework employs a DUM to refine these pseudo-labels.

(3) Effectiveness of the proposed two-branch framework. As shown in Figure 1, we establish a domain adaptation framework that includes two branches. The performance comparisons of different teacher-student frameworks are presented in Table 3. From the results in Table 3, we have the following findings: (a) Compared to ②, ③ and ④ obtain greater performance gains, indicating that adversarial learning is more effective than self-supervision. (b) When compared to ③, ④ does not consistently yield significant performance gains across all target domains, indicating that latent domains may not ensure the robustness of features for domain shifts. (c) In comparison with ②, ③, and ④, ⑤ and ⑥ achieve consistent performance improvements by exploring more branches to extract additional clues. (d) Compared to ⑤, our DUM helps generate more reliable pseudo-labels, significantly improving the accuracy in the target domain.

Comparison with state-of-the-art methods

To demonstrate the performance of our DBCL, we conduct numerous experiments in two aspects: intra-domain evaluation and cross-dataset evaluation.

(1) Intra-domain evaluation. FF++ dataset is used in

Method	Train	Test				Avg.	Δ
		DF	F2F	FS	NT		
Baseline [ICCV 19']	DF	-	65.2	42.5	68.3	58.7	-
FDA [CVPR 20']		-	62.3	55.4	58.4	58.7	+0.0
CORE [CVPR 22']		-	67.6	38.2	75.1	60.3	+1.6
DCL [AAAI 22']		-	77.1	**61.0**	75.0	71.0	+12.3
Ours		-	**87.6**	47.3	**87.3**	**74.1**	**+15.4**
Baseline [ICCV 19']	F2F	80.7	-	60.1	59.1	66.6	-
FDA [CVPR 20']		64.4	-	59.2	52.7	58.8	-7.8
CORE [CVPR 22']		73.9	-	53.6	49.0	58.8	-7.8
DCL [AAAI 22']		**91.9**	-	59.6	66.7	72.7	+6.1
Ours		83.8	-	**79.6**	**69.2**	**77.5**	**+10.9**
Baseline [ICCV 19']	FS	65.5	64.8	-	52.8	61.0	-
FDA [CVPR 20']		61.2	65.9	-	47.1	58.1	-2.9
CORE [CVPR 22']		66.1	66.3	-	**54.4**	62.3	+1.3
DCL [AAAI 22']		**74.8**	69.8	-	52.6	65.7	+4.7
Ours		72.1	**87.8**	-	52.8	**70.9**	**+9.9**
Baseline [ICCV 19']	NT	85.4	69.9	40.4	-	65.2	-
FDA [CVPR 20']		80.8	65.5	50.9	-	65.7	+0.5
CORE [CVPR 22']		83.6	69.3	47.4	-	66.8	+1.6
DCL [AAAI 22']		91.2	52.1	**79.3**	-	74.2	+9.0
Ours		**93.5**	**91.3**	76.0	-	**86.9**	**+21.7**

Table 4: Comparison with state-of-the-art methods w.r.t. AUC (%) on intra-domain evaluation. **Bold** indicates the best results, and underline indicates the second best results.

this experiment, where each sample from DF, F2F, FS, and NT is employed in its compressed form (HQ), as they are more challenging than the uncompressed versions. We train an Xception network (Chollet 2017) for binary classification as the baseline. In each experiment, we train the model on images generated by a single forgery method, and then calculate the AUC values (%) on images generated by other forgery methods. The results of our DBCL compared with other methods (i.e. FDA (Yang and Soatto 2020), CORE (Ni et al. 2022), and DCL (Sun et al. 2022)) are shown in Table 4. From the table, We have the following observations: (a) For all "source→target" pairs, our DBCL outperforms other methods w.r.t. the average AUC by a large margin, achieving state-of-the-art performance. (b) Compared to other methods, the improvements in our DBCL relative to the baseline are obvious. In particular, our DBCL brings an improvement of over 20% w.r.t AUC where NT serves as the source domain. (c) When we look into the AUC of each pair, our DBCL performs well across different pairs. It ranks 1st on 7 pairs, and ranks 2nd on 4 pairs.

We further analyze the feature space learned with the baseline and our DBCL. In Figure 5, we provide the t-SNE visualization (Liu et al. 2021b) of extracted features from Xception and our DBCL, respectively. It appears that our DBCL yields more discriminative domain-invariant features, which are successfully shrunk into the compact distribution (as shown in Figure 1).

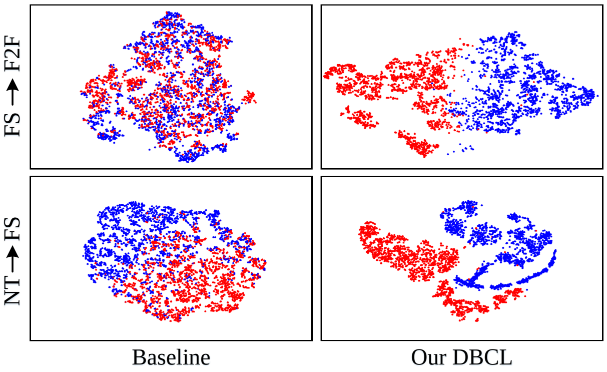

Figure 5: The t-SNE visualization of feature space. For each category (e.g. real/fake), the features are better aligned by our DBCL than baseline model (i.e. Xception).

Method	Celeb-DF-v2	ForgeryNet
Baseline [ICCV 19']	68.67	80.08
FDA [CVPR 20']	62.91	81.43
LipForensics [CVPR 21']	82.40	66.70
CORE [CVPR 22']	75.72	80.08
UIA-ViT [ECCV 22']	82.41	81.07
DCL [AAAI 22']	82.30	79.23
IID [CVPR 23']	83.80	81.35
Ours	**83.91**	**82.79**

Table 5: Comparison with state-of-the-art methods w.r.t. AUC (%) on cross-dataset evaluation. For all methods, FF++ (HQ) is used as the source dataset, and Celeb-DF-v2 and ForgeryNet are used as the target datasets.

(2) Cross-dataset evaluation. To further validate the generalization of our DBCL, we conduct cross-dataset evaluation using multiple datasets, including Celeb-DF-v2, FF++, and ForgeryNet. In the experiments, we employ FF++ (HQ) as the source dataset, Celeb-DF-v2 and ForgeryNet as the target datasets. We apply different methods (i.e. FDA (Yang and Soatto 2020), LipForensics (Haliassos et al. 2021), CORE (Ni et al. 2022), UIA-ViT (Zhuang et al. 2022), DCL (Sun et al. 2022), and IID (Huang et al. 2023a)) under the same experimental settings to compare their performances. The comparison results are illustrated in Table 5. It can be observed that our proposed DBCL outperforms other methods on both target datasets, demonstrating the effectiveness of our DBCL for cross-dataset deepfake detection.

CONCLUSION

In this paper, we propose a Dual-Branch Cooperative Learning (DBCL) method for domain adaptive deepfake detection. Within the proposed framework, domain alignment and knowledge distillation promote each other in a mutually reinforcing manner, yielding discriminative domain-invariant representations. The experimental results demonstrate the effectiveness of our DBCL for both intra-domain and cross-dataset deepfake detection. Considering the real-time application demands, in the future, we will focus on optimizing the speed of cross-domain deepfake detection methods.

References

Akkaya, I. B.; Altinel, F.; and Halici, U. 2021. Self-training guided adversarial domain adaptation for thermal imagery. In *CVPR Workshops*, 4322–4331.

Araslanov, N.; and Roth, S. 2021. Self-supervised augmentation consistency for adapting semantic segmentation. In *CVPR*, 15384–15394.

Chen, C.; Fu, Z.; Chen, Z.; Jin, S.; Cheng, Z.; Jin, X.; and Hua, X. 2020. HoMM: higher-order moment matching for unsupervised domain adaptation. In *AAAI*, 3422–3429.

Chen, L.; Zhang, Y.; Song, Y.; Liu, L.; and Wang, J. 2022. Self-supervised learning of adversarial example: towards good generalizations for deepfake detection. In *CVPR*, 18689–18698.

Chen, P.; Ye, J.; Chen, G.; Zhao, J.; and Heng, P. 2021a. Robustness of accuracy metric and its inspirations in learning with noisy labels. In *AAAI*, 11451–11461.

Chen, S.; Yao, T.; Chen, Y.; Ding, S.; Li, J.; and Ji, R. 2021b. Local relation learning for face forgery detection. In *AAAI*, 1081–1088.

Cheng, T.; Wang, X.; Chen, S.; Zhang, Q.; and Liu, W. 2023. BoxTeacher: exploring high-quality pseudo labels for weakly supervised instance segmentation. In *CVPR*, 3145–3154.

Chollet, F. 2017. Xception: deep learning with depthwise separable convolutions. In *CVPR*, 1800–1807.

Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; and Zafeiriou, S. 2020. RetinaFace: single-shot multi-level face localisation in the wild. In *CVPR*, 5202–5211.

Dong, S.; Wang, J.; Liang, J.; Fan, H.; and Ji, R. 2022. Explaining deepfake detection by analysing image matching. In *ECCV*, 18–35.

Fei, J.; Dai, Y.; Yu, P.; Shen, T.; Xia, Z.; and Weng, J. 2022. Learning second order local anomaly for general face forgery detection. In *CVPR*, 20238–20248.

Ganin, Y.; and Lempitsky, V. S. 2015. Unsupervised domain adaptation by backpropagation. In *ICML*, 1180–1189.

Gao, H.; Guo, J.; Wang, G.; and Zhang, Q. 2022. Cross-domain correlation distillation for unsupervised domain adaptation in nighttime semantic segmentation. In *CVPR*, 9903–9913.

Haliassos, A.; Mira, R.; Petridis, S.; and Pantic, M. 2022. Leveraging real talking faces via self-supervision for robust forgery detection. In *CVPR*, 14950–14962.

Haliassos, A.; Vougioukas, K.; Petridis, S.; and Pantic, M. 2021. Lips don't lie: a generalisable and robust approach to face forgery detection. In *CVPR*, 5039–5049.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

He, Y.; Gan, B.; Chen, S.; Zhou, Y.; Yin, G.; Song, L.; Sheng, L.; Shao, J.; and Liu, Z. 2021. ForgeryNet: a versatile benchmark for comprehensive forgery analysis. In *CVPR*, 4360–4369.

Hong, F.; Zhang, L.; Shen, L.; and Xu, D. 2022. Depth-aware generative adversarial network for talking head video generation. In *CVPR*, 3387–3396.

Huang, B.; Wang, Z.; Yang, J.; Ai, J.; Zou, Q.; Wang, Q.; and Ye, D. 2023a. Implicit identity driven deepfake face swapping detection. In *CVPR*, 4490–4499.

Huang, H.; Wang, Y.; Chen, Z.; Zhang, Y.; Li, Y.; Tang, Z.; Chu, W.; Chen, J.; Lin, W.; and Ma, K. 2022. CMUA-Watermark: a cross-model universal adversarial watermark for combating Deepfakes. In *AAAI*, 989–997.

Huang, L.; Li, Y.; Tian, H.; Yang, Y.; Li, X.; Deng, W.; and Ye, J. 2023b. Semi-supervised 2D human pose estimation driven by position inconsistency pseudo label correction module. In *CVPR*, 693–703.

Jeong, Y.; Kim, D.; Ro, Y.; and Choi, J. 2022. FrePGAN: robust deepfake detection using frequency-level perturbations. In *AAAI*, 1060–1068.

Kim, M.; Tariq, S.; and Woo, S. S. 2021. FReTAL: generalizing deepfake detection using knowledge distillation and representation learning. In *CVPR Workshops*, 1001–1012.

King, D. E. 2009. Dlib-ml: a machine learning toolkit. *J. Mach. Learn. Res.*, 10: 1755–1758.

Kumagai, A.; and Iwata, T. 2019. Unsupervised domain adaptation by matching distributions based on the maximum mean discrepancy via unilateral transformations. In *AAAI*, 4106–4113.

Kundu, J. N.; Kulkarni, A. R.; Bhambri, S.; Jampani, V.; and Radhakrishnan, V. B. 2022. Amplitude spectrum transformation for open compound domain adaptive semantic segmentation. In *AAAI*, 1220–1227.

Lai, Z.; Wang, C.; Cheung, S. S.; and Chuah, C. 2022. SaR: self-adaptive refinement on pseudo labels for multiclass-imbalanced semi-supervised learning. In *CVPR Workshops*, 4090–4099.

Lee, T.; Lee, B.; Shin, I.; Choe, J.; Shin, U.; Kweon, I. S.; and Yoon, K. 2022. UDA-COPE: unsupervised domain adaptation for category-level object pose estimation. In *CVPR*, 14871–14880.

Li, B.; Wang, Y.; Zhang, S.; Li, D.; Keutzer, K.; Darrell, T.; and Zhao, H. 2021a. Learning invariant representations and risks for semi-supervised domain adaptation. In *CVPR*, 1104–1113.

Li, J.; Xie, H.; Li, J.; Wang, Z.; and Zhang, Y. 2021b. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *CVPR*, 6458–6467.

Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; and Guo, B. 2020a. Face X-Ray for more general face forgery detection. In *CVPR*, 5000–5009.

Li, P.; Xu, Y.; Wei, Y.; and Yang, Y. 2022. Self-correction for human parsing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(6): 3260–3271.

Li, Y.; Yang, X.; Sun, P.; Qi, H.; and Lyu, S. 2020b. Celeb-DF: a large-scale challenging dataset for deepfake forensics. In *CVPR*, 3204–3213.

Liu, H.; Li, X.; Zhou, W.; Chen, Y.; He, Y.; Xue, H.; Zhang, W.; and Yu, N. 2021a. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *CVPR*, 772–781.

Liu, Y.; Zhang, S.; Li, Y.; and Yang, J. 2021b. Learning to adapt via latent domains for adaptive semantic segmentation. In *NeurIPS*, 1167–1178.

Liu, Y.; Zhang, S.; Yang, J.; and Yuen, P. C. 2021c. Hierarchical information passing based noise-tolerant hybrid learning for semi-supervised human parsing. In *AAAI*, 2207–2215.

Luo, Y.; Zhang, Y.; Yan, J.; and Liu, W. 2021. Generalizing face forgery detection with high-frequency features. In *CVPR*, 16317–16326.

Manjah, D.; Cacciarelli, D.; Standaert, B.; Benkedadra, M.; de Hertaing, G. R.; Macq, B.; Galland, S.; and De Vleeschouwer, C. 2023. Stream-based active distillation for scalable model deployment. In *CVPR Workshops*, 4998–5006.

Mazaheri, G.; and Roy-Chowdhury, A. K. 2022. Detection and localization of facial expression manipulations. In *WACV*, 2773–2783.

Melas-Kyriazi, L.; and Manrai, A. K. 2021. PixMatch: unsupervised domain adaptation via pixelwise consistency training. In *CVPR*, 12435–12445.

Mirza, M. J.; Micorek, J.; Possegger, H.; and Bischof, H. 2022. The norm must go on: dynamic unsupervised domain adaptation by normalization. In *CVPR*, 14745–14755.

Mittal, T.; Sinha, R.; Swaminathan, V.; Collomosse, J. P.; and Manocha, D. 2023. Video manipulations beyond faces: a dataset with human-machine analysis. In *WACV*, 643–652.

Ni, Y.; Meng, D.; Yu, C.; Quan, C.; Ren, D.; and Zhao, Y. 2022. CORE: consistent representation learning for face forgery detection. In *CVPR Workshops*, 12–21.

Nirkin, Y.; Wolf, L.; Keller, Y.; and Hassner, T. 2022. Deepfake detection based on discrepancies between faces and their context. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10): 6111–6121.

Nishi, K.; Ding, Y.; Rich, A.; and Höllerer, T. 2021. Improving label noise robustness with data augmentation and semi-supervised learning (student abstract). In *AAAI*, 15855–15856.

Petrovai, A.; and Nedevschi, S. 2022. Exploiting pseudo labels in a self-supervised learning framework for improved monocular depth estimation. In *CVPR*, 1568–1578.

Qu, X.; Zeng, J.; Liu, D.; Wang, Z.; Huai, B.; and Zhou, P. 2023. Distantly-supervised named entity recognition with adaptive teacher learning and fine-grained student ensemble. In *AAAI*, 13501–13509.

Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019. FaceForensics++: learning to detect manipulated facial images. In *ICCV*, 1–11.

Shiohara, K.; and Yamasaki, T. 2022. Detecting deepfakes with self-blended images. In *CVPR*, 18699–18708.

Sun, K.; Yao, T.; Chen, S.; Ding, S.; Li, J.; and Ji, R. 2022. Dual contrastive learning for general face forgery detection. In *AAAI*, 2316–2324.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 1195–1204.

Wang, X.; Huang, J.; Ma, S.; Nepal, S.; and Xu, C. 2022. Deepfake disrupter: the detector of deepfake is my friend. In *CVPR*, 14900–14909.

Wu, X.; Wu, Z.; Lu, Y.; Ju, L.; and Wang, S. 2022. Style mixing and patchwise prototypical matching for one-shot unsupervised domain adaptive semantic segmentation. In *AAAI*, 2740–2749.

Xu, R.; Yu, Y.; Cui, H.; Kan, X.; Zhu, Y.; Ho, J. C.; Zhang, C.; and Yang, C. 2023. Neighborhood-regularized self-training for learning with few labels. In *AAAI*, 10611–10619.

Xu, Y.; Yin, Y.; Jiang, L.; Wu, Q.; Zheng, C.; Loy, C. C.; Dai, B.; and Wu, W. 2022. TransEditor: transformer-based dual-space GAN for highly controllable facial editing. In *CVPR*, 7673–7682.

Yang, Y.; and Soatto, S. 2020. FDA: fourier domain adaptation for semantic segmentation. In *CVPR*, 4084–4094.

Ye, J.; Fu, C.; Zheng, G.; Paudel, D. P.; and Chen, G. 2022. Unsupervised domain adaptation for nighttime aerial tracking. In *CVPR*, 8886–8895.

Yeh, C.; Chen, H.; Shuai, H.; Yang, D.; and Chen, M. 2021. Attack as the best defense: nullifying image-to-image translation GANs via limit-aware adversarial attack. In *CVPR*, 16168–16177.

Yu, F.; Wang, D.; Chen, Y.; Karianakis, N.; Shen, T.; Yu, P.; Lymberopoulos, D.; Lu, S.; Shi, W.; and Chen, X. 2022. SC-UDA: style and content gaps aware unsupervised domain qdaptation for object detection. In *WACV*, 1061–1070.

Yu, F.; Zhang, M.; Dong, H.; Hu, S.; Dong, B.; and Zhang, L. 2021. DAST: unsupervised domain adaptation in semantic segmentation based on discriminator attention and self-training. In *AAAI*, 10754–10762.

Zhai, L.; Guo, Q.; Xie, X.; Ma, L.; Wang, Y. E.; and Liu, Y. 2022. A^{3}GAN: attribute-aware anonymization networks for face de-identification. In *ACM MM*, 5303–5313.

Zhang, D.; Lin, F.; Hua, Y.; Wang, P.; Zeng, D.; and Ge, S. 2022a. Deepfake video detection with spatiotemporal dropout transformer. In *ACM MM*, 5833–5841.

Zhang, J.; Huang, J.; Tian, Z.; and Lu, S. 2022b. Spectral unsupervised domain adaptation for visual recognition. In *CVPR*, 9819–9830.

Zheng, S.; Chen, C.; Yang, X.; and Tan, W. 2023. MaskBooster: end-to-end self-training for sparsely supervised instance segmentation. In *AAAI*, 3696–3704.

Zheng, Y.; Bao, J.; Chen, D.; Zeng, M.; and Wen, F. 2021. Exploring temporal coherence for more general video face forgery detection. In *ICCV*, 15024–15034.

Zhu, W.; Lu, L.; Xiao, J.; Han, M.; Luo, J.; and Harrison, A. P. 2022. Localized adversarial domain generalization. In *CVPR*, 7098–7108.

Zhuang, W.; Chu, Q.; Tan, Z.; Liu, Q.; Yuan, H.; Miao, C.; Luo, Z.; and Yu, N. 2022. UIA-ViT: unsupervised inconsistency-aware method based on vision transformer for face forgery detection. In *ECCV*, 391–407.